Transformer-based Pix2Pix GAN

Logan Garwood [1], Alex Albors Juez [1]

[1]Department of Mathematics, University of Washington

Introduction

- Pix2Pix is a popular GAN-based model used for image to image translation tasks.
- We consider different generator architectures, changing the classical U-net architecture for more modern approaches involving transformers.

Dataset

- Cityscapes is made of a large set of of high-quality pixel-level annotations. These consist of street scenes from 50 different cities. We train on 5000 images and test on 500. [1].
- Experimented with other datasets, available in the original Pix2Pix paper [3].

Architectures

- First, we used four transformer layers to encode a sequence of pixel patches.
- Then we introduce long residual connections and deepen the encoder to ten transformer layers.
- Finally, we use convolutions to embed patches into latent space before transforming the sequence.

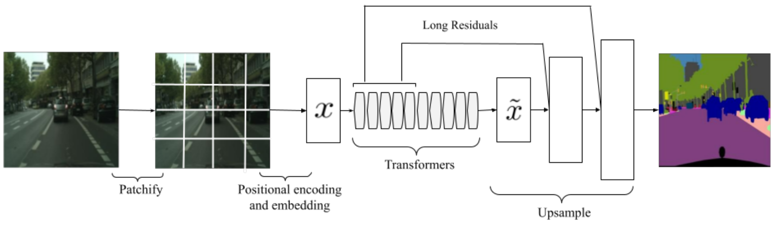

Figure 1. Architecture 2

Training

- We train for 500 epochs with Adam, with a fixed learning rate to 0.002, $\beta_1 = 0.5$ and $\beta_2 = 0.999$, as suggested in [3].

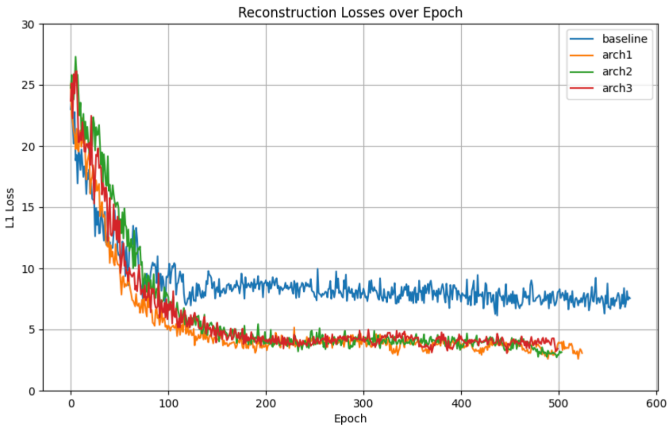

Figure 2. L_1 reconstruction loss throughout training

Conditional Generative Adversarial Network

In a conditional GAN, the discriminator and generator play a minimax game with objective function

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}\left[\log D(x, y)\right] + \mathbb{E}_{x}\left[\log(1 - D(x, G(x))\right]$$

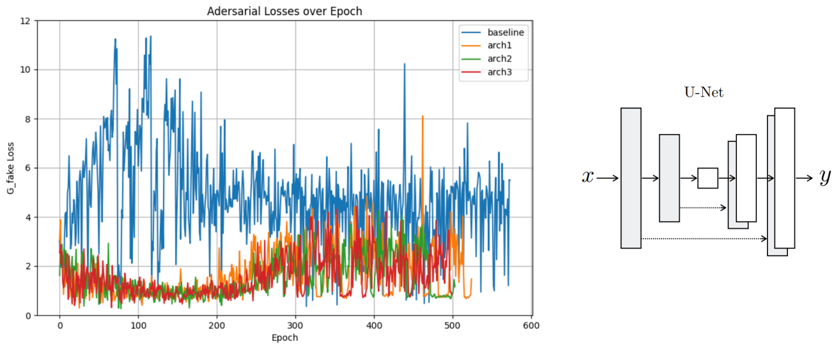

Figure 3. Generator losses and the classical U-Net architecture

Results

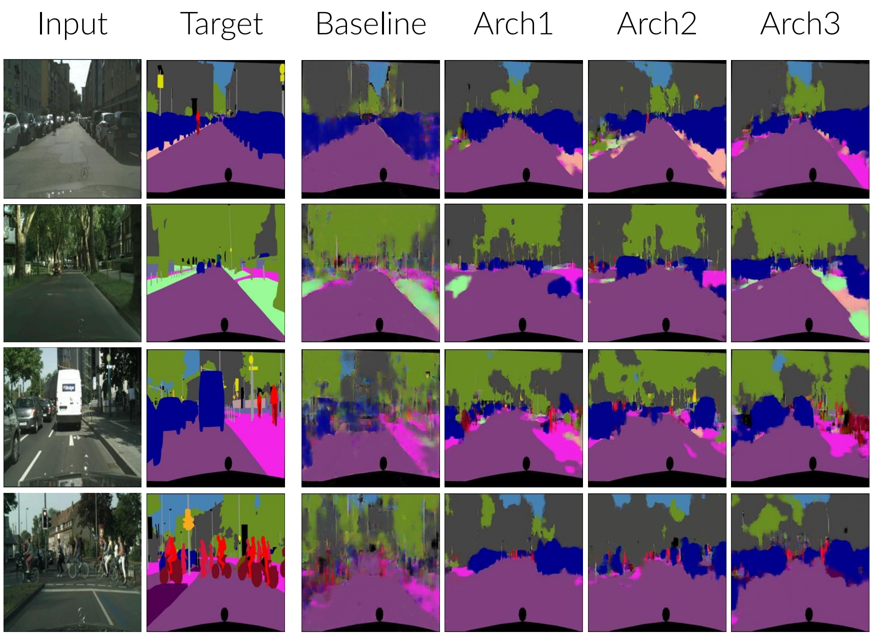

Table 1. Generated images on validation data

Model	FiD	Min Loss	Variance	Parameters (M)
Baseline	212.24	7.57	3.63	54.4
Arch1	130.39	3.30	0.91	23.9
Arch2	137.47	3.40	0.63	28.5
Arch3	134.67	3.90	0.64	28.9

Table 2. Comparison of Methods

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Scan the QR code to access more visual examples, datasets and all model implementations.